Issue Paper

RAND
EDUCATION

What Do Test Scores in Texas Tell Us?

Stephen P. Klein, Laura S. Hamilton, Daniel F. McCaffrey, Brian M. Stecher

Preface

During the past decade, several states have begun using the results on statewide tests as the basis for rewarding and sanctioning individual students, teachers, and schools. Although testing and accountability are intended to improve achievement and motivate staff and students, concerns have been raised in both the media and the professional literature (e.g., Heubert & Hauser, 1999; Linn, 2000) about possible unintended consequences of these programs.

The high-stakes testing program in Texas has received much of this attention in part because of the extraordinarily large gains the students in this state have made on its statewide achievement tests, the Texas Assessment of Academic Skills (TAAS). In fact, the gains in TAAS reading and math scores for both majority and minority students have been so dramatic that they have been dubbed the "Texas miracle." However, there are concerns that these gains were inflated or biased as an indirect consequence of the rewards and sanctions that are attached to the results. Thus, although there is general agreement that the gains on the TAAS are attributable to Texas' high-stakes accountability system, there is some question about what these gains mean. Specifically, do they reflect a real improvement in student achievement or something else?

We conducted several analyses to examine the issue of whether TAAS scores can be trusted to provide an accurate index of student skills and abilities. First, we used scores on the reading and math tests that are administered as part of the National Assessment of Educational Progress (NAEP) to investigate how much students in Texas have improved and whether this improvement is consistent with what has occurred nationwide. NAEP scores are a good benchmark for this purpose because they reflect national content standards and they are not subject to the same external pressures to boost scores as there are on the TAAS.

Next, we assessed whether the gains in TAAS scores between 1994 and 1998 were comparable to those on NAEP. We did this to examine how much confidence can be placed in the TAAS score gains. Similarly, we measured whether the differences in scores between whites and students of color on the TAAS were consistent with the differences between these groups on NAEP. Specifically, is the gap on TAAS credible given the gap on NAEP? And finally, we investigated whether TAAS scores are related to the scores on a set of three other tests that we administered to students in 20 Texas elementary schools.

Our findings from this research raise serious questions about the validity of the gains in TAAS scores. More generally, our results illustrate the danger of relying on statewide test scores as the sole measure of student achievement when these scores are used to make high-stakes decisions about teachers and schools as well as students. We anticipate that our findings will be of interest to local, state, and national educational policymakers, legislators, educators, and fellow researchers and measurement specialists.

Readers also may be interested in a RAND study by Grissmer et al. (2000) that compared the NAEP scores of different states across the country. Grissmer and his colleagues found that after controlling for various student demographic characteristics and other factors, Texas tended to have higher NAEP scores than other states and there was

some speculation as to whether this was due to the accountability system in Texas. Thus, while the Grissmer et al. (2000) report and the research presented in this issue paper both used NAEP scores, these studies differed in the questions they investigated, the data they analyzed, and the methodologies they employed. A forthcoming RAND issue paper will discuss some of the broader policy questions about high-stakes testing in schools.

The preparation of this issue paper benefited greatly from the many thoughtful suggestions and insights of our RAND colleagues, Dr. David Grissmer, Dr. Daniel Koretz, and Dr. James Thomson, and our external reviewers, Professor Richard Jaeger of the University of North Carolina at Greensboro and Professor Robert Linn of the University of Colorado at Boulder. We are also grateful to Miriam Polon and Christina Pitcher for editorial suggestions.

Background

Scores on achievement tests are increasingly being used to make decisions that have important consequences for examinees and others. Some of these "high-stakes" decisions are for individual students—such as for tracking, promotion, and graduation (Heubert & Hauser, 1999). Some states and school districts also are using test scores to make performance appraisal decisions for teachers and principals (e.g., merit pay and bonuses) and to hold schools and educational programs accountable for the success of their students (Linn, 2000). Although the policymakers who design and implement such systems often believe they lead to improved instruction, there is a growing body of evidence which indicates that high-stakes testing programs can also result in narrowing the curriculum and distorting scores (Koretz & Barron, 1998; Koretz et al., 1991; Linn, 2000; Linn, Graue, & Sanders, 1990; Stecher, Barron, Kaganoff, & Goodwin, 1998). Consequently, questions are being raised about the appropriateness of using test scores alone for making high-stakes decisions (Heubert & Hauser, 1999).

In this issue paper, we examine score gains on one statewide test in an effort to assess the degree to which they provide valid information about student achievement in that state and about improvements in achievement over time. This investigation is the latest in a decade-long series of RAND studies of high-stakes testing (e.g., Koretz & Barron, 1998). We believe that this work will provide lessons to help policymakers understand some of the challenges that arise in the context of high-stakes accountability systems.

Our interest in Texas was prompted by an unusual empirical relationship we observed between scores on TAAS and tests we administered to students in a small sample of schools as part of a larger study on teaching practices and student achievement. Because our set of schools was small and not representative of the state, we decided to explore statewide patterns of achievement on TAAS and on NAEP. In addition, Texas provides an ideal context in which to study high-stakes testing because its accountability system has received attention from the media and from the policy community, and it has been cited as possibly contributing to improved student achievement (e.g., Grissmer & Flanagan, 1998; Grissmer et al., 2000). TAAS scores are a central component of the accountability system. For example, students must pass the TAAS to graduate from high school, and TAAS scores affect performance evaluations (and, in some cases, compensation) for teachers and principals.

The TAAS program has been credited not only with improving student performance, but also with reducing differences in average scores among racial and ethnic groups. For example, a recent press release announced a record high passing rate on the TAAS. According to Commissioner of Education Jim Nelson, "Texas has justifiably gained national recognition for the performance gains being made by our students." Nelson also stated that Texas has "been able to close the gap in achievement between our minority youngsters and our majority youngsters, and we've again seen how we're progressing in that regard" (Jim Nelson as quoted by Mabin, 2000).

The unprecedented score gains on the TAAS have been referred to as the "Texas miracle." However, some educators and analysts (e.g., Haney, 2000) have raised questions about the validity of these gains and the possible negative consequences of high-stakes accountability systems, particularly for low-income and minority students. For example, the media have reported concerns about excessive teaching to the test, and there is some empirical support for these criticisms (Carnoy, Loeb, & Smith, 2000; McNeil & Valenzuela, 2000; Hoffman et al., in press). For instance, teachers in Texas say they are spending especially large amounts of class time on test preparation activities. Because the length of the school day is fixed, the more time that is spent on preparing students to do well on the TAAS often means there is less time to devote to other subjects.

There are also concerns that score trends may be biased by a variety of formal and informal policies and practices. For example, policies about student retention in grade may

affect score trends (McLaughlin, 2000). States may vary in the extent to which their schools promote students who fail to earn acceptable grades and/or statewide test scores. Eliminating these so-called "social promotions" would most likely raise the average scores at each *grade* level in subsequent years while lowering it at each *age* level. This is likely to occur because although the students who are held back may continue to improve, they are likely to do so at a slower rate than comparable students who graduate with their classmates (Heubert & Hauser, 1999). Another concern is inappropriate test preparation practices, including outright cheating. There have been documented cases of cheating across the nation, including in Texas. If widespread, these behaviors could substantially distort inferences from test score gains (Hoff, 2000; Johnston, 1999).

The pressure to raise scores may be felt most intensely in the lowest-scoring schools, which typically have large populations of low-income and minority students. Students at these schools may be particularly likely to suffer from overzealous efforts to raise scores. For example, Hoffman et al. (in press) found that teachers in low-performing schools reported greater frequency of test preparation than did teachers in higher-performing schools. This could lead to a superficial appearance that the gap between minority and majority students is narrowing when no change has actually occurred.

Evidence regarding the validity of score gains on the TAAS can be obtained by investigating the degree to which these gains are also present on other measures of these same general skills. Specifically, do the score trends on the TAAS correspond to those on the highly regarded NAEP? The NAEP tests are generally recognized as the "gold standard" for such comparisons because of the technical quality of the procedures that are used to develop, administer, and score these exams. Of course, NAEP is not a perfect measure. For example, there are no stakes attached to NAEP scores, and therefore student motivation may differ on NAEP and state tests, such as TAAS. However, it is currently the best indicator available.

There are several other reasons why score gains on the TAAS are not likely to have a one-to-one match with those on NAEP if these tests assess different skills and knowledge. However, the specifications for the NAEP exams are based on a consensus of a national panel of experts, including educators, about what students should know and be able to do. Hence, NAEP provides an appropriate benchmark for measuring improvement. As Linn (2000) notes, "Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement" (p. 14).

Questions for Our Research

Understanding the source and consequences of the impressive score gains on the TAAS would require an extensive independent study. We have not done that. Instead, the analyses described below address the following questions about student achievement in Texas:

1. Have the reading and math skills of Texas students *improved* since the full statewide implementation of the TAAS program in 1994 (e.g., are fourth graders reading better today than fourth graders a few years ago); and, if their skills did improve: (a) how much improvement occurred and (b) was the amount of improvement in reading the same as it was in math?
2. Are the gains in reading and math on the TAAS consistent with what would be expected given NAEP scores in Texas and the rest of the country?
3. Has Texas *narrowed the gap* in average reading and math skills between whites and students of color?
4. Do other tests given in Texas at a sample of 20 schools produce results that are consistent with those obtained with the TAAS?

We begin by describing certain important features of the TAAS and NAEP exams. We then answer the first three questions through analyses of publicly available TAAS and NAEP data and discuss the findings. Next, we answer the fourth question by reporting the results from a study that administered other tests to about 2,000 Texas students. Finally, we present our conclusions.

Description of the TAAS

TAAS was initiated in 1990 to serve as a criterion-referenced measure of the state's mandated curriculum. It is intended to be comprehensive and to measure higher-order thinking skills and problem-solving ability (Texas Education Agency, 1999). Since the full implementation of the TAAS program in 1994, it has been administered in reading and mathematics in grades 3, 4, 5, 6, 7, 8, and 10. Other subjects are also tested at selected grade levels. Last

year, for example, a writing test was given at grades 4, 8, and 10. Science and social studies were tested at grade 8. The TAAS tests consist primarily of multiple-choice items, but the writing test includes questions that require written answers.

Teachers administer the TAAS tests to their own students. Answers are scored by the state. The questions are released to the public after each administration of the exam, and a new set of TAAS tests is administered each year. However, the format and content of the questions in one year are very similar to those used the next year. Each form of the TAAS contains items that are being field-tested for inclusion in the forms to be used in subsequent years. These items are also used to link test scores from one year to the next to help ensure consistent difficulty over time. These experimental items are not used to compute student scores nor are they released to the public. This practice is consistent with that employed in many other large-scale testing programs.

The TAAS is administered only in Texas. Thus, there are no national norms or benchmarks against which to compare the performance of Texas students on this test. However, the Texas Education Agency administered the Metropolitan Achievement Tests to a sample of Texas students to determine how well these students performed relative to a national norm group. We discuss this study in a later section of this issue paper.

Description of NAEP

The national portion of NAEP is mandated by Congress and is administered through the National Center for Education Statistics. It is currently the only assessment that provides information on the knowledge and skills of a representative sample of the nation's students. The content of NAEP tests is based on test specifications that were developed by educators and others, and is intended to reflect a consensus about what students should be learning at a given grade level. Hence, the questions are not tied to standards of a single state or district.[1] Like TAAS, NAEP is designed to assess problem-solving skills in addition to content knowledge. A national probability sample of schools is invited to participate in NAEP. Schools that decline are replaced with schools where the student characteristics are similar to those at the schools that refused to participate.

Most states, including Texas, also arrange to have the NAEP exams administered to another (and larger) group of their schools to allow for the generation of reliable state-level results. This state-level testing utilizes the same general procedures as the national NAEP program does; e.g., third-party selection of the participating schools and having a cadre of trained consultants (rather than classroom teachers) administer the tests. However, unlike the national program, these consultants may be local district personnel.

In both the national and state-level programs, a given student is asked a sample of all the questions that are used at that student's grade level. This permits a much larger sampling of the content domain in the available testing time than would be feasible if every student had to answer every item. Different item formats (including multiple-choice, short-answer, and essay) are used in most subjects. The breadth of content and item types, as well as the consensus of a national panel of experts that is reflected in NAEP frameworks, makes NAEP a useful indicator of achievement trends across the country.

The validity of NAEP scores is enhanced by the procedures that are used to give the exams and ensure test security (e.g., test administrators do not have a stake in the outcomes). However, the utility of NAEP scores is limited by some of the other features of this testing program. For instance, NAEP is not administered every year, and when it is administered, not every subject is included, only a few grade levels are tested, and individual student, school, and district scores are not available. These features preclude examining year-to-year trends in a particular subject or tracking individual student progress over time. The motivation to do well on the NAEP tests is intrinsic rather than driven by external stakes. However, any reduction in student effort or performance that may stem from NAEP being a relatively low-stakes test should be fairly consistent over time and therefore not bias our measurement of score improvements across years.

How We Report Results

NAEP and TAAS results are typically reported to the public in terms of the percentage of students passing or meeting certain performance levels (or "cut" scores). Although this type of reporting seems easier to understand, it can lead to

[1] It was beyond the scope of this issue paper to identify the specific similarities and differences in content coverage between NAEP and TAAS.

erroneous conclusions. For example, the difficulty of achieving a passing status or a certain level of performance (such as "proficient") may vary between tests as well as within a testing program over time. Making comparisons based on percentages reaching certain levels also does not account for score changes among students who perform well above or below the cut score.

To avoid these and other problems with percentages, we adopted the research community's convention of reporting results in terms of "effect" sizes. The effect size is the difference in mean scores (between years or groups) divided by the standard deviation of those scores. In other words, it is the standardized mean difference. The major advantage of using effect sizes is that they provide a common metric across tests.

As a frame of reference for readers who are not familiar with this metric, the effect size for the difference in achievement between white and black students has ranged from 0.8 to 1.2 across a variety of large-scale tests (Hedges & Nowell, 1998). The effect size for the difference in third grade student reading scores between large and small classes in Tennessee was approximately 0.25 (Finn & Achilles, 1999).[2]

Have Reading and Math Skills Improved in Texas?

NAEP data have been cited as evidence of the effectiveness of educational programs in Texas (e.g., Grissmer & Flanagan, 1998). For instance, within a racial or ethnic group, the average performance of the Texas students tends to be about six percentile-points higher than the national average for that group (Grissmer et al., 2000; Reese et al., 1997).

These results are consistent with the findings obtained by the Texas Education Agency in its 1999 Texas National Comparative Data Study, in which a sample of Texas students took the Metropolitan Achievement Tests, Seventh Edition (MAT-7). Texas students at every grade level scored slightly higher than the national norming sample in most subjects (Texas Education Agency, 1999). However, it is difficult to draw conclusions from this study because, according to the sampling plan for this research, each participating school selected the classrooms and students that would take the MAT. Moreover, Texas did not report the mean TAAS scores of the students who took the MAT. Under the circumstances, the TAAS data are vital for determining whether those who took the MAT were truly representative of their school or the state. For example, the interpretation of the MAT findings would no doubt change if it was discovered that the mean TAAS scores of the students who took the MAT were higher than the corresponding state mean TAAS scores.

Data from a single year cannot tell us whether achievement has improved over time or whether trends in TAAS scores are reflected in other tests. To answer the question of whether performance improved, we compared the scores of Texas fourth graders in one year with the scores of Texas fourth graders four years later. We did this in both reading and mathematics. We also did this for eighth graders in mathematics (NAEP's testing schedule precluded conducting a similar analysis for eighth graders in reading). We then contrasted these results with national trends to assess whether the gains in Texas after the full statewide implementation of the TAAS differed from those in other states.

The main finding is that the average test score gains on the NAEP in Texas exceeded those of the nation in only one of the three comparisons.

Figures 1 through 3 present the results of these analyses. The main finding is that over a four-year period, the average test score gains on the NAEP in Texas exceeded those of the nation in only one of the three comparisons, namely: fourth grade math.

Figure 1 shows that the Texas fourth graders in 1998 had higher NAEP reading scores than did Texas fourth graders in 1994. The size of the increase was .13 standard deviation units for white students and .15 units for students of color. However, these increases were not unique to Texas. The national trend was for all students to improve. In fact, only among white fourth graders was the improvement in Texas greater than improvement nationally, and then only slightly (the difference in the effect sizes between

[2]This estimate includes students who spent one to four years in small classes.

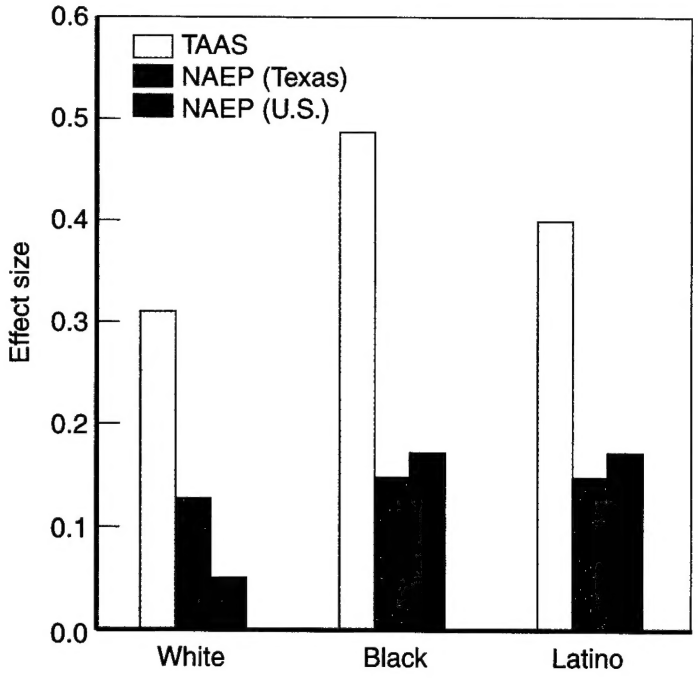

Figure 1
Reading Effect Sizes for 4th Graders on NAEP and TAAS Between 1994 and 1998

Texas and the United States was .08). We discuss the implications of this difference in score gains between groups when we discuss the question of whether Texas has narrowed the gap in performance among racial and ethnic groups.

The TAAS data tell a radically different story (see Figure 1). They indicate there was a very large improvement in TAAS reading scores for all groups (effect sizes ranged from .31 to .49). Figure 1 also shows that on the TAAS, black and Hispanic students improved more than whites. The gains on TAAS were therefore several times larger than they were on NAEP. And, contrary to the NAEP findings, the gains on TAAS were greater for students of color than they were for whites.

Figure 2 shows that fourth graders in Texas in 1996 had substantially higher NAEP math scores than did fourth graders in 1992 (effect sizes ranged from .25 to .43). Moreover, this improvement was substantially greater than the increase nationwide. This was especially true for white stu-

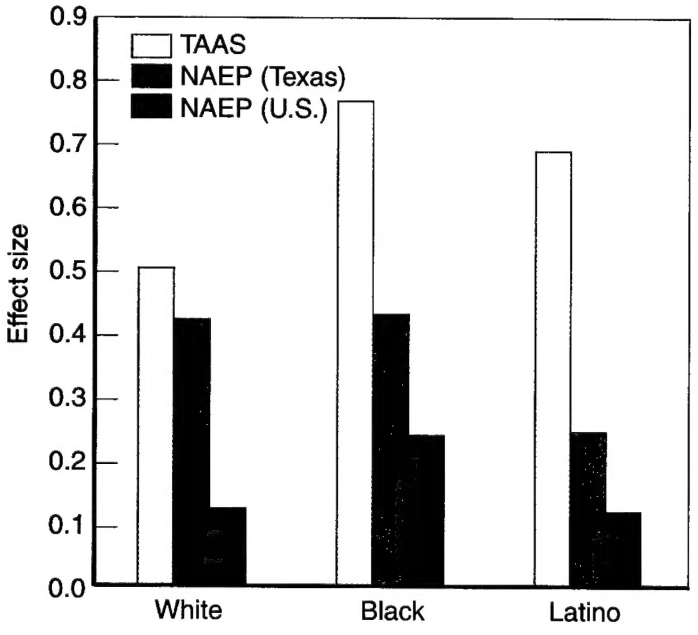

Figure 2
Math Effect Sizes for 4th Graders on NAEP Between 1992 and 1996, and on TAAS Between 1994 and 1998

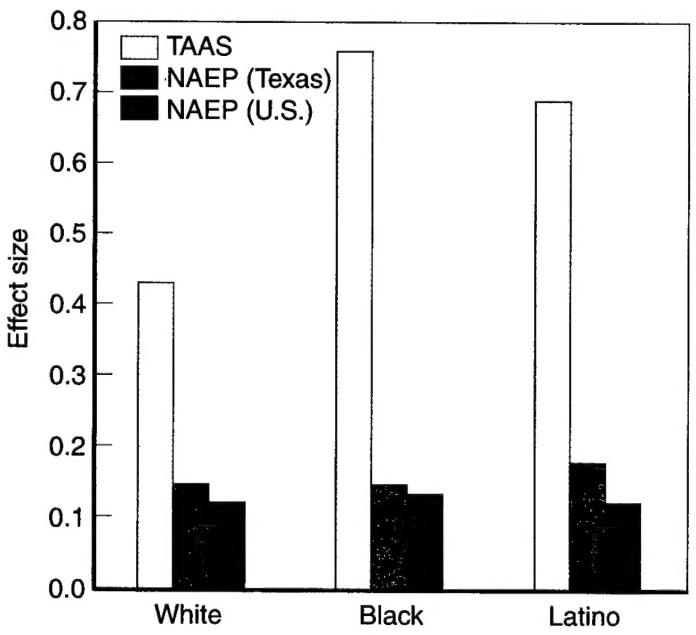

Figure 3
Math Effect Sizes for 8th Graders on NAEP Between 1992 and 1996, and on TAAS Between 1994 and 1998

dents. Nevertheless, the gains on TAAS were much larger than they were on NAEP, especially for students of color.[3]

Figure 3 shows that Texas eighth graders in 1996 had higher NAEP scores than did Texas eighth graders in 1992, but these differences were only slightly larger than those observed nationally. Thus, as with fourth grade reading, there was nothing remarkable about the NAEP scores in Texas, and students of color did not gain more than whites. In contrast, there were huge improvements in eighth grade math scores on the TAAS during a similar four-year period, and these increases were much larger for students of color than they were for whites. The same was true for eighth grade TAAS reading scores during this period (effect sizes for whites, blacks, and Hispanics were .28, .45, and .37, respectively).

To further examine the question of whether there has been an improvement in reading and math skills of Texas students, we compared the NAEP scores of fourth graders in one year with the NAEP scores of eighth graders four years later. Because of the way NAEP samples students for testing, this is analogous (but not equivalent) to following the same cohort of students over time. In fact, the redesign of NAEP in 1984, which established a practice of testing grade levels four years apart and conducting the assessment in the core subjects every four years, was intended in part to support this type of analysis (Barton & Coley, 1998). We present results for Texas and the nation so readers can see the extent to which Texas students are progressing relative to students in other states.

Table 1 shows that the average NAEP math scale score for white Texas fourth graders in 1992 was 229. Four years later, the mean score for white eighth graders was 285, i.e., a 56-point improvement. However, there was a 54-point improvement nationally for whites during this same period. There was a similar pattern for minority students, and these trends held for both math and reading (Table 2). In short, the score increases in Texas were almost identical to those nationwide (we could not conduct the corresponding analysis with TAAS data because TAAS does not convert scores to a common scale across grade levels).

[3]In Figures 2 and 3, the NAEP and TAAS trends cover different but overlapping years, due to the testing schedules of these measures.

Is Texas Closing the Gap Between Whites and Students of Color?

In 1998, the mean fourth grade NAEP reading score for whites in Texas was one full standard deviation higher than the mean for blacks. To put this in perspective, the average black student was at roughly the 38th percentile among all Texas test takers whereas the average white student was at about the 67th percentile. This gap was slightly larger than the difference between these groups in 1994. In other words, the black-white reading gap actually increased during this four-year period. The same pattern was present in fourth and eighth grade math scores (see Figure 4a).

In contrast, the difference in mean TAAS scores between whites and blacks was initially smaller than it was on NAEP, and it decreased substantially over a comparable four-year period. Consequently, by 1998, the black-white gap on TAAS was about half what it was on NAEP. In other words, whereas the gap on NAEP was large to begin with and got slightly wider over time, the gap on TAAS started off somewhat smaller than it was on NAEP and then got substantially smaller.

Table 1
Mean NAEP Math Scores for 4th Graders in 1992 and 8th Graders in 1996

	Texas			United States			Texas-U.S.
Group	4th	8th	Gain	4th	8th	Gain	
White	229	285	56	227	281	54	2
Black	199	249	50	192	242	50	0
Hispanic	209	256	47	201	250	49	–2

Table 2
Mean NAEP Reading Scores for 4th Graders in 1994 and 8th Graders in 1998

	Texas			United States			Texas-U.S.
Group	4th	8th	Gain	4th	8th	Gain	
White	227	273	46	223	270	47	–1
Black	191	245	54	186	241	55	–1
Hispanic	198	252	54	188	243	55	–1

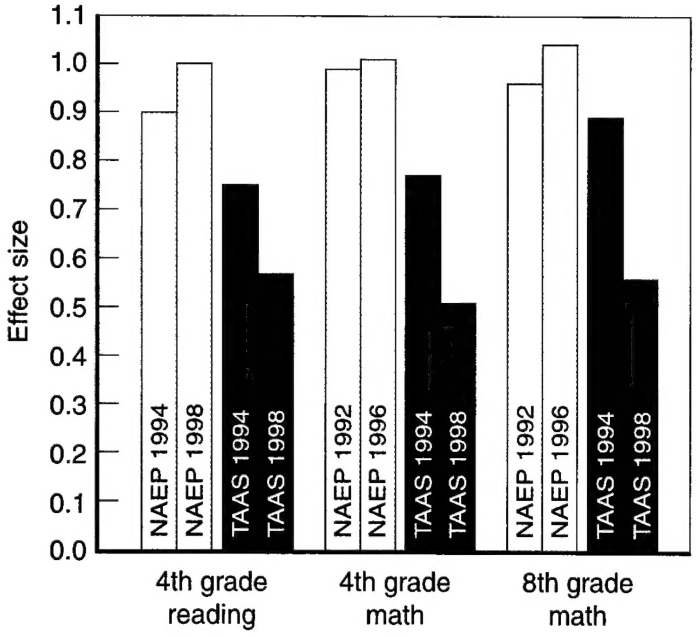

Figure 4a
Difference Between White and Black Mean Scores in Texas on NAEP and TAAS (in Standard Deviation Units)

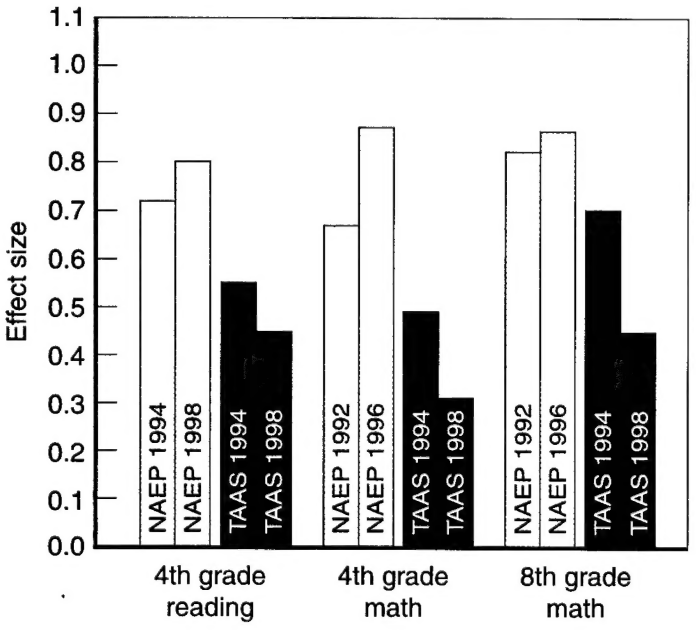

Figure 4b
Difference Between White and Hispanic Mean Scores in Texas on NAEP and TAAS (in Standard Deviation Units)

The same radically disparate NAEP and TAAS trends were also present for the Hispanic-white gap; i.e., the gap got slightly wider on NAEP but substantially smaller on TAAS over comparable four-year periods (see Figure 4b). In addition, although fourth grade math was the subject on which Texas showed the largest gains over time relative to the nation, the white-Hispanic NAEP gap grew in Texas but not nationally, and the white-black gap remained constant in Texas but actually shrank nationally. In short, gap sizes on NAEP were moving in the opposite direction than they were on TAAS.

It is worth noting that even the relatively small NAEP gains we observed might be somewhat inflated by changes in who takes the test. As mentioned earlier, Haney (2000) provides evidence that exclusion of students with disabilities increased in Texas while decreasing in the nation, and Texas also showed an increase over time in the percentage of students dropping out of school and being held back. All of these factors would have the effect of producing a gain in average test scores that overestimates actual changes in student performance.

Why Do TAAS and NAEP Scores Behave So Differently?

The large discrepancies between TAAS and NAEP results raise serious questions about the validity of the TAAS scores. We do not know the sources of these differences. However, one plausible explanation, and one that is consistent with some of the survey and observation results cited earlier, is that many schools are devoting a great deal of class time to highly specific TAAS preparation. It is also plausible that the schools with relatively large percentages of minority and poor students may be doing this more than other schools.

TAAS questions are released after each administration. Although there is a new version of the exam each year, one version looks a lot like another in terms of the types of questions asked, terminology and graphics used, content areas covered, etc. Thus, giving students instruction and practice on how to answer the specific types of questions that appear on the TAAS could very well improve their scores on this exam. For example, in an effort to improve their TAAS scores, some schools have retained outside contractors to work with teachers, students, or both.

If the discrepancies we observed between NAEP and TAAS were due to some type of focused test preparation for

the TAAS, then this instruction must have had a fairly narrow scope. With the possible exception of fourth grade math, it certainly did not appear to influence NAEP scores. In short, if TAAS scores were affected by test preparation for the TAAS, then the effects of this preparation did not appear to generalize to the NAEP exams. This explanation also raises questions about the appropriateness of what is being taught to prepare students to take the TAAS.

A small but significant percentage of students may have "topped out" on the TAAS. In other words, their TAAS scores may not reflect just how much more proficient they are in reading and math than are other students. If that happened, it would artificially narrow the gap on the TAAS between whites and students of color (because majority students tend to earn higher scores than minority students). Thus, the reduced gap on the TAAS relative to NAEP may be an artifact of the TAAS being too easy for some students.[4] If so, it also would deflate the gains in TAAS scores over time. In short, were it not for any topping-out, the TAAS gain scores in Figures 1 through 3 would have been even larger, which in turn would further increase the disparity between TAAS and NAEP results.

What Happens on Other Tests?

We collected data on about 2,000 fifth graders from a mix of 20 urban and suburban schools in Texas. This study was part of a much larger project that included administering different types of science and math tests to students who also took their state's exams. The 20 schools were from one part of Texas. They were not selected to be representative of this region let alone of Texas as a whole. Nevertheless, some of the results at these schools also raised questions about the validity of the TAAS as a measure of student achievement.

[4]The results in the 20-school study discussed later in this issue paper suggest that some topping-out occurred on the TAAS. For example, although about two-thirds of the 2,000 students in this study were in a free or reduced-price lunch program, 7 percent answered 95 percent of the TAAS reading questions correctly and 9 percent did so on the math test. Only a few students were able to do this on any of the tests we gave.

Test Administration

In the spring of 1997, our Texas students took the English language version of the TAAS in reading and math. A few weeks later, we administered the following three tests to these same students: the Stanford 9 multiple-choice science test, the Stanford 9 open-ended (OE) math test, and a "hands-on" (HO) science test developed by RAND (Stecher & Klein, 1996). The Stanford 9 OE math test asked students to construct their own answers and write them in their test booklets. In the HO science test, students used various materials to conduct experiments. They then wrote their answers to several open-ended questions about these experiments in a simulated laboratory notebook. Table 3 shows the means and standard deviations on each measure.

Some Expected and Unexpected Findings

We analyzed the data in two ways. First, we investigated whether the students who earned high scores on one test tended to earn high scores on the other tests. Next, we examined whether the schools that had a high average score on one test tended to have high average scores on the other tests. We also looked at whether the results were related to type of test used (i.e., multiple-choice or open-ended), subject matter tested (reading, math, or science), and whether a student was in a free or reduced-price school lunch program. The latter variable serves as a rough indicator of a

Table 3
Means and Standard Deviations on Supplemental Study Measures by Unit of Analysis

	Students		Schools	
Variable	Mean	Standard Deviation	Mean	Standard Deviation
TAAS math	37.97	13.62	38.84	3.80
TAAS reading	29.33	10.61	29.61	2.59
Stanford 9 science	29.01	5.40	28.55	1.94
Stanford 9 OE math	15.14	5.21	14.84	1.44
HO science	11.78	6.00	11.44	1.83
Percentage in lunch program (SES)	67.84	46.7	76.10	22.3

NOTES: TAAS math had 52 items and TAAS reading had 40 items. Stanford 9 science had 40 items. The maximum possible scores on Stanford 9 OE math and HO science were 27 and 30, respectively.

student's socioeconomic status (SES). For the school-level analyses, SES was indicated by the percentage of students at the school who were in the subsidized lunch program.

Some of our results were consistent with those in previous studies. Others were not. We begin with what was consistent and then turn to those that were anomalous.

The first column of Table 4 shows the correlation between various pairs of measures when the student (N ≈ 2,000) is the unit of analysis.[5] The second column shows the results when the school (N = 20) is the unit of analysis. The first set of rows show that the measures we administered correlated about .55 with each other when the student was the unit of analysis. These correlations were substantially higher when the school was the unit. For example, the correlation between Stanford 9 science and Stanford 9 OE math was .55 when the student was the unit, but it was .78 when the school was the unit. These results are very consistent with the general findings of other research on student achievement.

Table 4
Correlations Between Measures

Correlations between:	Unit of analysis	
	Students	Schools
Non-TAAS tests		
Stanford 9 science and HO science	.57	.88
Stanford 9 science and Stanford 9 OE math	.55	.78
Stanford 9 OE math and HO science	.53	.71
SES and non-TAAS tests		
SES and Stanford 9 science	−.17	−.76
SES and Stanford 9 OE math	−.10	−.72
SES and HO science	−.18	−.66
SES and TAAS tests		
SES and TAAS math	−.08	.13
SES and TAAS reading	−.14	−.21
TAAS and non-TAAS tests		
TAAS math and Stanford 9 science	.48	−.07
TAAS math and Stanford 9 OE math	.46	.02
TAAS math and HO science	.48	.03
TAAS reading and Stanford 9 science	.52	.10
TAAS reading and Stanford 9 OE math	.42	.21
TAAS reading and HO science	.53	.13
TAAS math and TAAS reading	.81	.85

The second set of rows in Table 4 shows a strong negative correlation between the percentage of students at a school who were in the lunch program and that school's mean on the tests we administered. In other words, schools with more affluent students tended to earn higher mean scores on the non-TAAS tests than did schools with less wealthy students. This relationship is present regardless of test type (multiple-choice or open-ended) and subject matter (math or science). Again, these findings are very consistent with those found in other testing programs.

The correlation between SES and our test scores is much stronger when the school is used as the unit of analysis than when the student is the unit. This is a common finding and stems in part from the fact that it is difficult to get a high correlation with a dichotomous variable (i.e., in program versus not in program). The school-level analyses do not suffer from this problem because SES at the school level is measured by the percentage of students at the school who are in the program (i.e., a continuous rather than a dichotomous variable). School-level analyses also tend to produce higher correlations than individual-level analyses because aggregation of scores to the school level reduces the percentage of error in the estimates.

The anomalies appear in the third and fourth sets of rows. In the third set, SES had an unusually small (Pearson) correlation with both of the TAAS scores even when the

[5]The correlation coefficient, which can range from −1.00 to +1.00, is a measure of the degree of agreement between two tests. A high positive correlation is obtained when the students (or schools) that have high scores on one test also tend to have high scores on the other test.

school was used as the unit of analysis.[6] This result (which is opposite to the one we found with the non-TAAS tests) was due to a curvilinear relationship between SES and TAAS scores. Specifically, schools with a relatively low or high percentage of students in the lunch program tended to have higher mean TAAS math scores than did schools with an average percentage of students in this program (see Figure 5). Thus, the typical relationship between SES and test scores disappeared on the TAAS even though this relationship was present on the tests we administered a few weeks after the students took the TAAS. Figure 6 illustrates the more typical pattern by showing the negative, linear relationship between Stanford 9 math test scores and the percentage of students in the free or reduced-price lunch program.

The fourth set of rows in Table 4 shows that when the student is the unit of analysis, TAAS math and reading scores correlate well with the scores on the tests we gave. Although the correlations are somewhat lower than would be expected from experience with other tests (especially the .46 correlation between the two math tests), these differences do not affect the conclusions we would make about the relationships among different tests. However, the correlation between TAAS and non-TAAS tests essentially disappears when the school is the unit of analysis. This result is contrary to the one that would be expected by other studies and the results in the first block of rows.

The last row of Table 4 shows that TAAS math has a very high correlation with TAAS reading (despite being a different subject). In fact, TAAS math correlates much higher with TAAS reading than it does with another math test (namely: Stanford 9 OE math).

To sum up, the non-TAAS tests correlated highly with each other and with SES; and, as expected, this correlation increased when the school was used as the unit of analysis. Also as anticipated, the two TAAS tests had a moderate correlation with the non-TAAS tests, but unexpectedly, this only occurred when the student was used as the unit of analysis. Rather than getting larger, the correlation between TAAS and non-TAAS tests essentially evaporated when the school was the unit. And finally, regardless of the unit of

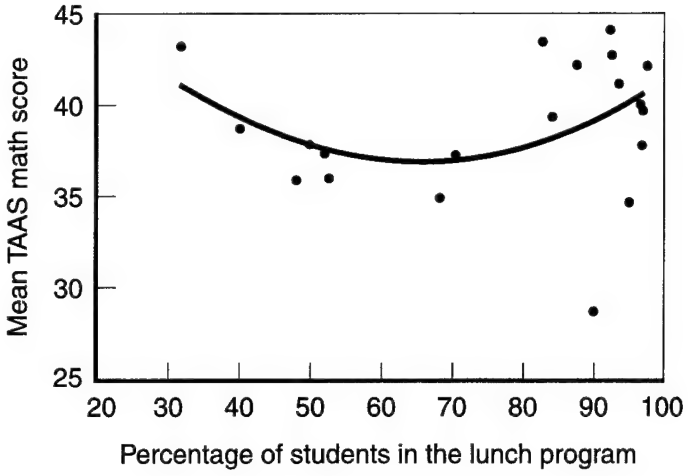

Figure 5
Relationship Between Mean TAAS Math Score and Percentage of Students in the Lunch Program When the School Is Used As the Unit of Analysis

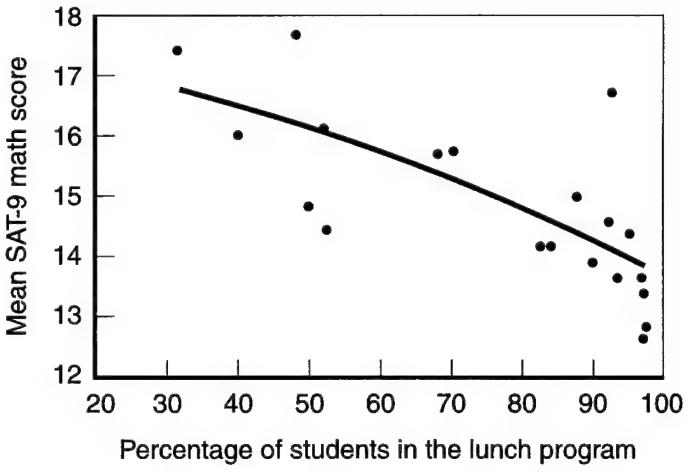

Figure 6
Relationship Between Mean SAT-9 Math Score and Percentage of Students in the Lunch Program When the School Is Used As the Unit of Analysis

[6]We also examined the relationships by splitting the schools into two groups, according to whether they had relatively high versus low percentages of students in the lunch program (e.g., those that had more than 70 percent versus those with less than 70 percent). This analysis produced results that were consistent with the data in Figures 5 and 6. Specifically, schools with a high percentage of students in the lunch program had much lower scores on the three tests we gave than did schools with a relatively low percentage of students in this program whereas that was not the case with the TAAS scores.

analysis, the two TAAS tests had an extremely high correlation with each other, but both had a virtually zero correlation with SES.

One of the reasons we were surprised that the TAAS and non-TAAS scores behaved so differently is that the latter tests were designed to measure some of the same kinds of higher-order thinking skills that the TAAS is intended to measure. However, our results could be due to the unique characteristics of the 20 schools in our study or other factors. We are therefore reluctant to draw conclusions from our findings with these schools or to imply that these findings are likely to occur elsewhere in Texas. Nevertheless, they do suggest the desirability of periodic administration of external tests to validate TAAS results. This procedure, which is sometimes referred to as "audit testing," could have been incorporated into the study of the Metropolitan Achievement Test discussed previously.

Conclusions

We are now ready to answer the questions that we posed at the beginning of this issue paper. Specifically, we found that the reading and math skills of Texas students improved since the full implementation of the TAAS program in 1994. However, the answers to the questions of how much improvement occurred, whether the improvement in reading was comparable to what it was in math, and whether Texas reduced the gap in scores among racial and ethnic groups depend on whether you believe the NAEP or TAAS results. They tell very different stories.

NAEP and TAAS results tell us very different stories.

According to NAEP, Texas fourth graders were slightly more proficient in reading in 1998 than they were in 1994. However, the country as a whole also improved to about the same degree. Thus, there was nothing remarkable about reading score gains in Texas. In contrast, the increase in fourth grade math scores in Texas was significantly greater than it was nationwide. However, the small improvements in NAEP eighth grade math scores were consistent with those observed nationally. The gains in scores between fourth and eighth grade in Texas also were consistent with national trends. In short, except for fourth grade math, the gains in Texas were comparable to those experienced nationwide during this time period.

In all the analyses, including fourth grade math, the gains on the TAAS were several times greater than they were on NAEP. Hence, how much a Texas student's proficiency in reading and math actually improved depends almost entirely on whether the assessment of that student's skills relies on NAEP scores (which are based on national content frameworks) or TAAS scores (which are based on tests that are aligned with Texas' own content standards and are administered by the classroom teacher).

The huge disparities between the stories told by NAEP and TAAS are especially striking in the assessment of (1) the size of the gap in average scores between whites and students of color and (2) whether these gaps are getting larger or smaller. According to NAEP, the gap is large and increasing slightly. According to TAAS, the gap is much smaller and decreasing greatly. We again quote Linn (2000, p. 14): "Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement." Put simply, how different could "reading" and "math" be in Texas than they are in the rest of the country?

The data available for this report were not ideal. Limitations in the way NAEP is administered make it difficult to do the kinds of comparisons that would be most informative. For example, NAEP is not given every year and individual student or school scores are not available. And the supplemental study described above was limited to 20 schools in just one part of a very large state. Nevertheless, the stark differences between TAAS and NAEP (and other non-TAAS tests) raise very serious questions about the generalizability of the TAAS scores.

These concerns about TAAS do not condemn all efforts to increase accountability, nor should they be interpreted as being opposed to testing. On the contrary, we believe that some form of large-scale assessment, when properly implemented, is an essential tool to monitor student progress and thereby support state efforts to improve education. Moreover, the possible problems with the TAAS discussed earlier in this issue paper are probably not restricted to this test or state. For example, score inflation and unwanted test preparation have been found in a number of jurisdictions (Koretz

& Barron, 1998; Linn, 2000; Stecher et al., 1998; Heubert & Hauser, 1999).

To sum up, states that use high-stakes exams may encounter a plethora of problems that would undermine the interpretation of the scores obtained. Some of these problems include the following: (1) students being coached to develop skills that are unique to the specific types of questions that are asked on the statewide exam (i.e., as distinct from what is generally meant by reading, math, or the other subjects tested); (2) narrowing the curriculum to improve scores on the state exam at the expense of other important skills and subjects that are not tested; (3) an increase in the prevalence of activities that substantially reduce the validity of the scores; and (4) results being biased by various features of the testing program (e.g., if a significant percentage of students top out or bottom out on the test, it may produce results that suggest that the gap among racial and ethnic groups is closing when no such change is occurring).

There are a number of strategies that states might try to lessen the risk of inflated and misleading gains in scores. They can reduce the pressure to "raise scores at any cost" by using one set of measures to make decisions about individual students and another set (employing sampling and third-party administration) to make decisions about teachers, schools, and educational programs. States can replace their traditional paper-and-pencil multiple-choice exams with computer based "adaptive" tests that are tailored to each student's abilities, that draw on "banks" of thousands of questions, and that are delivered over the Internet into the school building (for details, see Bennett, 1998; Hamilton, Klein, & Lorie, 2000). States can also periodically conduct audit testing to validate score gains. They can study the positive and negative effects of the testing program on curriculum and instruction, and whether these effects are similar for different groups of students. For instance, what knowledge, skills, and abilities are and are not being developed when the focus is concentrated on preparing students to do well on a particular statewide, high-stakes exam? However, given the findings reported above for Texas, it is evident that something needs to be done to ensure that high-stakes testing programs, such as the TAAS, produce results that merit public confidence and thereby provide a sound basis for educational policy decisions.

References

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade* (ETS Policy Information Report). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (1998). *Reinventing assessment.* Princeton, NJ: Educational Testing Service.

Carnoy, M., Loeb, S., & Smith, T. L. (2000). *Do higher state test scores in Texas make for better high school outcomes?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21,* 97–109.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us.* Santa Monica, CA: RAND, MR-924-EDU.

Hamilton, L. S., Klein, S. P., & Lorie, W. (2000). *Using web-based testing for large-scale assessment.* Santa Monica, CA: RAND, IP-196-EDU.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8* (41). Available at http://epaa.asu.edu/epaa/v8n41.

Hedges, L. V., & Nowell, A. (1998). Black-white test score convergence since 1965. In Jencks, C., & Phillips, M. (Eds.), *The Black-White Test Score Gap* (pp. 149–181). Washington, DC: Brookings.

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation.* A Report of the National Research Council, Washington, DC: National Academy Press.

Hoff, D. J. (2000). As stakes rise, definition of cheating blurs. *Education Week,* June 21.

Hoffman, J. V., Assaf, L., Pennington, J., & Paris, S. G. (in press). High stakes testing in reading: Today in Texas, tomorrow? *The Reading Teacher.*

Johnston, R. C. (1999). Texas presses districts in alleged test-tampering cases. *Education Week,* March 17.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND, MR-1014-EDU.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests, in R. L. Linn (chair), *The Effects of High Stakes Testing,* symposium presented at the annual meetings of the American Educational

Research Association and the National Council on Measurement in Education, Chicago, April.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4–16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9,* 5–14.

Mabin, Connie (2000). State's students again improve on TAAS scores. *Austin American-Statesman,* May 18.

McLaughlin, D. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates.* Palo Alto, CA: American Institutes for Research.

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric.* Cambridge, MA: Harvard University Civil Rights Project.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 report card for the nation and the states.* Washington, DC: National Center for Education Statistics.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing* (CSE Technical Report 482). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., & Klein, S. P. (Eds.) (1996). *Performance assessments in science: Hands-on tasks and scoring guides.* Santa Monica, CA: RAND, MR-660-NSF.

Texas Education Agency (1999). *Texas Student Assessment Program: Technical digest for the academic year 1998–1999.* Available at http://www.tea.state.tx.us/student.assessment/techdig.htm.

Texas Education Agency (2000). *1999 Texas National Comparative Data Study.* Available at http://www.tea.state.tx.us/student.assessment/researchers.htm.

Texas Education Agency (2000). *Texas TAAS passing rates hit seven-year high; four out of every five students pass exam.* Press release, May 17.

Building on more than 25 years of research and evaluation work, RAND Education has as its mission the improvement of educational policy and practice in formal and informal settings from early childhood on.

RAND

1700 Main Street, P.O. Box 2138, Santa Monica, California 90407-2138 • Telephone 310-393-0411 • FAX 310-393-4818
1200 South Hayes Street, Arlington, Virginia 22202-5050 • Telephone 703-413-1100 • FAX 703-413-8111

IP-202 (2000)